DOCUMENT RESUME

ED 121 823 TM 005 254

AUTHOR Rossi, Robert Joseph
TITLE A Philosophical Perspective on Construct Validation: Application of Inductive Logic to the Analysis of Experimental Episode Construct Validity.
PUB DATE [Apr 76]
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS Achievement Need; *Inductive Methods; Information Theory; *Logic; Probability; Projective Tests; Statistical Analysis; *Test Validity
IDENTIFIERS *Experimental Episode Construct Validity

ABSTRACT

Methods drawn from four logical theories associated with studies of inductive processes are applied to the assessment and evaluation of experimental episode construct validity. It is shown that this application provides for estimates of episode informativeness with respect to the person examined in terms of the construct and to the construct definition itself. Example analyses are presented in which projective and objective tests of the achievement motive are evaluated in relation to a version of the theory of achievement motivation presented by Atkinson (1957). Results indicate why projective measures may be preferred. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A PHILOSOPHICAL PERSPECTIVE ON CONSTRUCT VALIDATION: APPLICATION OF INDUCTIVE LOGIC TO THE ANALYSIS OF EXPERIMENTAL EPISODE CONSTRUCT VALIDITY*

Robert Joseph Rossi

American Institutes for Research

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

Robert J. Rossi

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

ED121823

TM005 254

Introduction

The process of construct validation is an important concern of both educational and psychological researchers. Lord and Novick (1968), for example, have referred to construct validity as the most important characteristic of a test for scientific purposes, while Campbell and Stanley (1963) have referred to the interpretability and representativeness of experiments as minima for understanding experimental results. A clear statement of what is the logic of construct validation was presented by Cronbach and Meehl (1955) in an earlier paper entitled "Construct Validity in Psychological Tests." Their remarks offer an appropriate introduction to the present study, although concern here is primarily for demonstrating the applicability of certain logical theories for analysis of this validity. These authors wrote:

> A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct. We expect a person at any time to possess or not possess a qualitative attribute (amnesia) or structure, or to possess some degree of a quantitative attribute (cheerfulness). A construct has certain associated meanings carried in statements of this general character: Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized or loose, used in

*This paper is adapted from the author's dissertation entitled Toward a Theory of Construct Validity: An Application of Inductive Logic to the Analysis of Experimental Episodes and is the complete version of a paper presented at the 1976 annual meeting of the American Educational Research Association in San Francisco.

AERA Session Number 26.11

> ramified theory or in a few simple propositions, used in absolute propositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test... (p. 247)

A test or an experiment is therefore valid with respect to a construct when it produces results that can be interpreted in terms of the construct definition under consideration.

Tests and experiments that are construct valid provide researchers with two types of information: information about the persons tested in terms of the construct (e.g., "S/he is highly motivated to achieve") and information about the construct definition itself--how it can be strengthened, extended, and so on. These two ways in which a construct valid test, for example, can be informative offer some perspective on how the construct validity of tests generally may be understood. High construct validity signals that a test is likely to be highly informative with respect to both persons tested and the construct definition being considered, whereas lower degrees of this validity may suggest high person-low construct informativeness, low person-high construct informativeness, or low person-low construct informativeness. Depending on the researcher's primary purpose, selection of an appropriately construct valid test may be made by review of how much information of what type(s) alternative tests can provide. Measurement of the type and amount of information associated with particular degrees of construct validity is thus important for both the assessment and the evaluation of this experimental episode characteristic.

It is unfortunate that presently used (statistical) methods for assessing and evaluating construct validity neither utilize nor are related to formal measures of information. Since the assumption usually made is that high degrees of validity are preferable in every situation, correlation coefficients suffice as indicators of those tests that are best. However, while this approach is feasible when the choice is between valid and less than valid tests,[1] it falls short in less than extreme cases. For these cases a way of evaluating validity that is related to informativeness would be of use. This paper is thus concerned with the evaluation of test and experimental

---

[1] This choice is, perhaps, characterized best within the sort of matrix proposed by Campbell and Fiske (1959). In this matrix both correlations with tests designed to assess the same construct and those with tests designed to assess different constructs are presented and may be compared.

episode construct validity in terms of the type and amount of information that can be provided by these episodes. The approach taken is mainly a logical rather than an empirical one and should therefore be regarded as a preliminary step toward explicating logically what is meant by saying of a test or experiment that it is construct valid. Unlike the logical characterization of this validity offered by Tuomela (1970), however, the one presented here does consider directly the relation between models of experiments (experimental episodes) and theoretical models. As a result, this application of inductive methods to experimental episode analysis should provide both philosophers and researchers with a new perspective on current research practices.

The measurement of information provided by experimental episodes requires clarity with respect to the kind of information that is to be assessed. In a report entitled A Generalized Psychometric Theory Based on Information Measure, Cronbach (1952) attempted to use the mathematical theory of communication developed by Shannon and Weaver (1949) to characterize aspects of tests and testing in terms of their informativeness. However, as Bar-Hillel pointed out in 1952, this mathematical theory is not concerned with the meaning of transmitted signals but only with their frequency. Thus, since the kind of information Cronbach was interested in was the meaningful information provided by tests and experiments,[2] it is not surprising that in a later

---

[2]This interest is evidenced by his use of the term "information" in passages such as the following:

> The choice between tests may be based on either the amount of true information offered or the certainty of whatever is given. (Less attention is given to exhaustiveness because, for any given criterion, J is a function of I.) Sometimes the message that is extremely trustworthy (i.e., all it says is true) is relatively unsatisfactory because it says so little. Associated Press news reports have generally been regarded as very valid. Columnists and commentators whose information is less saturated with truth enjoy a large audience, however, because they report so many more things than AP. Among their gossip, rumor, and speculation, enough is true that readers or listeners value it, even though each item must be taken *cum grano salis*. In hiring employees, letters of recommendation are notoriously fallible, yet nearly every personnel man seeks them because they give so much more information than the mere statement that "X worked for us 2-1/2 years as a shipping clerk." The latter information is more dependable, but less complete. [emphasis added] (p. 8)

In each instance reference was clearly being made to the meaningfulness of the information provided.

paper (Cronbach, 1953) he appraised critically his earlier effort. Since it is the meaningful information provided by tests and experiments that must be assessed in order to evaluate degrees of test or experiment validity, it is a semantic information theory that is needed.

Such a theory of information was proposed in 1952 by Carnap and Bar-Hillel in a technical report entitled An Outline of a Theory of Semantic Information. Based on Carnap's theory of logical probability, this information theory shares some formulas with the Shannon and Weaver theory but, unlike this mathematical interpretation, is aimed at assessing the amount of meaningful information conveyed by a statement to one who understands it. Use of the Carnap and Bar-Hillel theory to assess the informativeness of tests and experiments was suggested by these authors themselves in concluding their Outline but, as yet, has not been attempted. This is somewhat surprising insofar as the theory of semantic information has within it the means for describing both the type and the amount of meaningful information provided by experimental episodes. In order to describe how this semantic theory can be practically applied, however, it is necessary to review briefly the theories of distributive normal forms (Hintikka, 1953, 1965), confirmation (Carnap, 1950), semantic information, and definability (Hintikka and Tuomela, 1970).

A. Distributive normal forms in first-order logic

First-order logic or quantification theory is a language which contains the sentential connectives (and, or, not, if...then, if and only if), free individual terms, quantifiers and variables bound to them, and predicates describing either attributes of individuals or relations between them. In this study only one-place (monadic) predicates are used but methods and results may be generalized to more complex cases. Of principal importance here is the description of attributive constituents (a-constituents) and constituents, key elements in the derivation of normal forms. These elements are best described in terms of a simplified example.

Let the one-place predicates A and B have meaning as follows:

A = is male

B = is a philosopher

If these predicates are conjoined and are permuted with respect to negation, four kinds of individuals can be described:

$AB(x)$ = X is a male philosopher

$\sim AB(x)$ = X is a female philosopher

$A\sim B(x)$ = X is a male non-philosopher

$\sim A\sim B(x)$ = X is a female non-philosopher

In general, given κ monadic predicates $2^{\kappa} = K$ kinds of individual descriptions can be formed. Each such description is called an a-constituent. By means of these a-constituents a given universe of individuals can be partitioned into pairwise exclusive and mutually exhaustive classes or kinds. Constituents or logical descriptions of kinds of possible worlds are formed from a-constituents in the sense that constituents or possible worlds tell which kinds of individuals exist and which do not in a particular universe. Thus, from the four a-constituents above $2^4 = 16$ descriptions of possible kinds of worlds can be formed, one of which is the following:

$$(Ex)\ AB(x)\ \&\ (Ex)\ \sim AB(x)\ \&\ (x)\ (AB(x)\ v\ \sim AB(x))$$

This statement describes a world in which the only kinds of individuals are male philosophers and female philosophers.

For any statement S, then, it is possible to list the possible kinds of worlds in which S can be true. The disjunction of those worlds in which S can be true is called the (disjunctive) normal form of S. Distributive normal forms are simply disjunctive normal forms of S occurring at different depths; that is, when S is either generalized or further specified. In the present study the disjunctive normal forms of statements will include only constituents containing a single layer of quantifiers. It is in terms of normal forms of this type that inductive probability or confirmation theory may be understood most easily.

B. Confirmation theory

While statistical probability can be understood as a quantitative characteristic of physical systems established empirically, logical probability is

of a purely logical nature requiring no empirical tests. This should not suggest, however, that this probability is unrelated to empirical matters. Carnap (1951) wrote:

> A statement of inductive probability states a relation between a hypothesis and a given body of evidence, e.g., results of actual or possible observations. The asse ted probability value means the degree to which the hypothesis is confirmed or supported by the evidence. It is important to notice that a statement on inductive probability or degree of confirmation is relative to the evidence. (p. 20)

What is at issue in the study of logical probability or confirmation theory is the relation between two types of statements--statements of inductive generalization or singular inductive inference and those of evidence. In particular, concern is for establishing the degree to which what is meant by an inductive statement is confirmed or supported by some statement of evidence. Confirmation theory is thus rooted in the meaningfulness of statements.

Generally speaking, the a priori and a posteriori (logical) probability values of generalizations and singular hypotheses are determined as soon as probability values are assigned to those constituents and a-constituents respectively in their normal forms and to those constituents and a-constituents in the normal forms of related evidential statements. Hintikka (1966) has suggested the use of a two-parameter system for making these assignments, adding to Carnap's λ-continuum an "α-dimension" concerned expressly with inductive generalization. The parameters λ and α are quasi-logical terms, values for which are determined by logical considerations together with other knowledge that is relevant to the particular inductive situation. α and λ can be thought of either as indices of caution with respect to inductive generalization and inductive inference respectively or as the means by which Bayesian prior probabilities are assigned to constituents and a-constituents. The larger their values the more caution must one show in making any inductive statement. An intuitive way to regard α is as an a priori estimate of the number of different kinds of individuals in the universe over which generalizations are to be made. Similarly λ may be thought of as an a priori estimate of the size of a sample of individuals to be divided equally among those a-constituents which occur as members of a constituent.

On the basis of α and λ values and the structures of the particular constituents and a-constituents being considered, formulas presented by Hintikka may be used to assign a priori and a posteriori probability values to a-constituents, constituents, and, on the basis of these assignments, statements with these a-constituents and constituents in their normal forms. The components of these formulas and the formulas themselves are presented below.

1. α, λ, K, w. α and λ are terms which assign Bayesian prior probabilities to constituents and a-constituents. Their values lead to different inductive dispositions; that is, they affect the ways in which generalizations and singular inferences are made on the basis of evidence. K is the total number of logically possible a-constituents that can be formed within a particular universe. (In the example presented earlier, K = 16.) w stands for the "width" or number of a-constituents included in a particular constituent. The larger is w, the more encompassing is the constituent.

2. e, c, n, n_j. e refers to evidence consisting of the observation of exactly n individuals exemplifying c a-constituents. n and c thus define e and also define n_j. n_j is just the number of individuals observed that exemplify the particular a-constituent j so that $\sum_{j=1}^{c} n_j = n$.

3. Cw, h. Cw refers to a particular constituent consisting of w a-constituents. Since it describes a possible kind of world in which there are only w kinds of individuals, it may be regarded as a "strong generalization" in terms of which still more general statements will be assigned probability values. h refers to an hypothesis advanced concerning a particular a-constituent within Cw. Unlike Cw which is a generalization, h describes a singular inference from the data.

The formulas in which these terms occur are the following. Note that the term π refers to the product of two terms, e.g., $\pi(X,Y) = Y\cdot(1+Y)\cdot(2+Y)\cdot\ldots\cdot(X-1+Y)$.

$$(1) \qquad p(C_w) = \pi\left(\alpha, \frac{w\lambda}{K}\right) \Big/ \sum_{i=0}^{K} \binom{K}{i} \pi\left(\alpha, \frac{i\lambda}{K}\right)$$

$$(2) \qquad p(C_w|e) = \frac{\pi(\alpha, w\lambda/K) \prod_{j=1}^{c} \pi(n_j, \lambda/w)}{\sum_{i=0}^{K-c} \left[\binom{K-c}{i} \pi(\alpha, (c+i)\lambda/K) \prod_{j=1}^{c} \pi(n_j, \lambda/(c+i))\right]}$$

$$(3) \qquad p(e) = \frac{\sum_{i=0}^{K-c} \binom{K-c}{i} \pi(\alpha, (c+i)\lambda/K) \prod_{j=1}^{c} \pi(n_j, \lambda/(c+i))}{\sum_{i=0}^{K} \left[\binom{K}{i} \pi(\alpha, i\lambda/K)\right] \cdot \pi(n, \lambda)}$$

$$(4) \qquad p(h|e) = \frac{\sum_{i=0}^{K-c} \left[\binom{K-c}{i} \pi(\alpha, (c+i)\lambda/K) \prod_{j=1}^{c} \pi(n_j, \lambda/(c+i))(n_1 + \lambda/(c+i))\right]}{\sum_{i=0}^{K-c} \left[\binom{K-c}{i} \pi(\alpha, (c+i)\lambda/K) \prod_{j=1}^{c} \pi(n_j, \lambda/(c+i))\right](n+\lambda)}.$$

In order to bring out the significance of α and λ for determining probability values, Table 1 presents the results of two example analyses based on these formulas. In the first example α and λ are each set equal to 2 whereas in the second example they are each set equal to 200. All other terms are defined in the same ways for both cases: K=32, w=6, c=6, n=20, n_1=5, n_2=4, n_3=3, n_4=3, n_5=3, and n_6=2. Differences in the resulting probability values are thus due to the different values of these two parameters.

TABLE 1. Example Analyses Based on Formulas (1)-(4)

	α=2, λ=2	α=200, λ=200
p(Cw)	$.6 \times 10^{-10}$	$.14 \times 10^{-75}$
p(Cw/e)	$.13 \times 10^{-4}$	$.26 \times 10^{-62}$
p(e)	$.23 \times 10^{-23}$	$.13 \times 10^{-28}$
p(h/e)	.23	$.5 \times 10^{-1}$

C. Semantic information theory

The basic idea underlying the concept of information first formalized by Carnap and Bar-Hillel in the early fifties is that the information conveyed by a statement S is equivalent to the amount of uncertainty one is relieved of when one learns that S is true. In order to measure this uncertainty, distinction must be made between those elements (constituents and a-constituents) in the normal form of S and those not included. The more

possible kinds of worlds or possible kinds of individuals S includes or "leaves open" the less it can be said to reduce one's uncertainty concerning the world; that is, the more possibilities with respect to the world must one prepare for. Therefore, the more possibilities S excludes the fewer possibilities must one prepare for, and the more informative is S. Since the probabilities of a-constituents and constituents in the normal form of a statement determine the probability of the statement itself, the probability of S and its informativeness are inversely related.

This inverse relationship between probability and informativeness has played a prominent role in Karl Popper's conception of science and the scientific search for truth. In his paper "Degree of Confirmation," Popper (1954) wrote:

> *Science does not aim, primarily, at high probabilities. It aims at a high informative content, well backed by experience. But a hypothesis may be very probable simply because it tells us nothing, or very little.* A high degree of probability is therefore not an indication of "goodness"--it may be merely a symptom of low informative content. (p. 146)

Carnap and Bar-Hillel formally represented this relationship with the following basic formulas:

$$\text{cont}(S) = 1 - p(S)$$

$$\text{inf}(S) = -\log p(S)$$

"Cont" is often referred to as a measure of the substantive information or content of S while "inf" can be thought of as a measure of the unexpectedness of (the truth of) S or, more simply, as a measure of its surprise value. Both measures are required to express adequately what is meant by the information conveyed by a statement; that is, one sometimes must rely on cont rather than inf to describe how informative is some S (and vice versa). The formal relation of these measures is straightforward, $\text{inf}(S) = \log 1/1 - \text{cont}(S)$. An example provided by Bar-Hillel in his paper entitled "Semantic Information and Its Measures" should help to clarify this relation conceptually. Drawing on Agatha Christie's title Bridge Murder Case as inspiration, he described the following scenario:

> There was a bridge party in A's villa, with B, C, D, and E participating; A was the host and only kibitzed. When the last rubber was finished and the guests were looking for A to take leave of him, they found him murdered in the garden. Every one of the four players had been the dummy at one time or another and had left the room for refreshments. Each one had, on the available evidence, an equal opportunity for murdering A. A reward was promised to those who could forward information leading to the identification of the murderer.
>
> A day later, X came and produced evidence sufficient to prove that B could not have been the murderer. The next day Y showed, to the district attorney's satisfaction, that C was innocent. The following day, Z did the same for D. Whereupon E was duly convicted and electrocuted.
>
> The problem now for the district attorney was how the reward should be distributed; he had to adopt some numerical proportion. He could evaluate the information given by the three informants according to the absolute cont value of their statements, or according to their absolute inf value, or according to their measures, relative to the information he received, or according to any explicit function whatsoever. (pp. 307-308)

Using a cont function, the district attorney might distribute the reward unequally in favor of X, while with an inf function the unequal distribution should favor Z. The difference lies in whether it is the overall reduction in uncertainty (or substantial increase in knowledge) each individual's evidence makes possible (cont) or whether it is the greater unexpectedness of the truth revealed by this evidence that is to be more highly prized.

In addition to the absolute cont and inf measures presented above, several relative measures have been proposed (see Hintikka, 1967). Two of these relative measures of informativeness to be used here are conditional content ($cont_{cond}$) and conditional information (inf_{cond}). These are defined as follows:

$$cont_{cond}\ (S/T) = 1 - p\ (S/T)$$
$$inf_{cond}\ (S/T) = -\log\ (S/T)$$

These measures assess the informativeness of a statement S in a world where the state of affairs described by T exists. Relative measures of this kind are especially important for evaluating the substantive content and the surprise value of a-constituents, constituents, and more encompassing statements with these elements in their normal forms on the basis of evidence (e.g., Cw/e, h/e). To be useful for evaluating degrees of construct validity in

terms of person and construct informativeness, however, these measures must be "supplemented" by one other logical theory: the theory of definability.

Since the ways in which the parameters α and λ are defined determine the manner in which probability values are assigned to constituents and a-constituents, α and λ determine the absolute and relative informativeness of statements with these constituents and a-constituents in their normal forms. If, therefore, the definition of these quasi-logical parameters is based on the interpretive relation between an experimental episode and some construct of interest, then probability values will reflect how likely are statements made on the basis of episode results and in terms of the construct to be true. Measures of information will thus refer to the amount of uncertainty these statements remove regarding the construct definition itself. More specifically, probability values derived from application of the four formulas listed above will, in this case, serve as indicators of person informativeness (i.e., how likely this statement describing this individual in terms of this construct is to be true), while absolute and relative cont and inf measures will serve as indicators of construct informativeness (e.g., how much has this statement increased one's knowledge of this construct). The exact manner in which these parameters are to be defined so as to reflect the interpretive relation between experimental episode and construct definition is outlined in a later section of this paper. Definability theory, which is necessary for estimating degrees of interpretability, is discussed here as it makes use of logical elements already reviewed.

D. Definability theory

One way of making operational the notion of interpretability is to regard it in terms of the degree to which elements from one list define particular elements on another list. If, for example, the "first list" is comprised of individual test item descriptions of a person tested and the "second list" is the definition of some construct of interest, interpretability of test results in terms of the construct will be in proportion to the degree to which test item descriptions can define the construct. Definability theory aims at describing this relation between terms describing observable phenomena and theoretical terms defined within nomological networks.

For example, suppose that a theory of X-ness (a construct definition) consists of the following four axioms:

1. $A \,\&\, B \vee C \rightarrow X$
2. $\neg B \vee \neg C \rightarrow A \,\&\, X$
3. $A \,\&\, \neg C \vee \neg B \rightarrow \neg X$
4. $B \,\&\, C \rightarrow A$

A, B, and C are observational terms or predicates and X, the construct in this case, is a theoretical term. From the combination of these four predicates $2^4=16$ a-constituents can be formed in the manner described earlier. Similarly, from these 16 a-constituents $2^{16}=65,536$ constituents are formed. Each constituent, describing as it does a possible kind of world, may be considered as a possible model for the four-axiom theory. To determine if a constituent <u>is</u> a model for this theory one must see whether the kinds of individuals it includes satisfy each of the four axioms; that is, whether the values of predicates described by each of its "individuals" can be substituted into each of the axioms with the result being a logically true statement. Evaluating the ways in which X is defined by A, B, and C in each model of the theory thus provides a measure of the interpretability of these A, B, and C descriptions in terms of X.

Hintikka and Tuomela have outlined three ways in which such definition can occur: explicitly, conditionally, and partially. Due to the preliminary nature of the present effort, however, consideration is given to explicit definability only. Thus, for every model of the four-axiom theory described, the same set of a-constituents must define X and in no model can an a-constituent define both X and $\neg X$. X is said, then, to be defined by the set of a-constituents associated with this theoretical predicate in every model of the theory and not defined otherwise. In the example there is only one constituent that can serve as a model for the theory presented and this constituent contains the single a-constituent ABCX. Therefore the predicate combination ABC defines X in this case, and, should an experimental episode describe a person tested in terms of the predicates ABC, one would be justified in saying that the person is (or has) X. In a later section, where a simplified version of the theory of achievement motivation is considered and compared to two test episodes designed to measure the achievement motive, the procedure described above is followed and results are used

to provide definitions for α and λ so that the person and construct informativeness of these test episodes can be evaluated.

E. Construct, external, internal validities

The definitions of α and λ to be presented later in this paper are not governed by logical considerations alone. Both the issues of experimental episode interpretability and representativeness play major roles in the use of logical formulas. To understand why these issues are so important to the present study it is necessary to understand clearly how construct, external, and internal validities can be related in cases where generalization is to be made from experimental episode to theory and inference to future, similarly designed episodes must be made in terms of this episode-theory relationship.

External validity has to do with the overall "representativeness" of experimental episodes (see Snow, 1974). Representativeness refers to the similarities between the episode itself and the universe of generalization. Thus an externally valid experiment is one which allows for generalizations to be made to a universe or to several universes of interest with relatively little difficulty. This validity is assessed in two ways: first, a qualitative comparison is made between the episode and the universe of generalization, and second, an assessment is made of the interpretability of those factors within an experiment that can restrict results to the particular experimental context. As Campbell and Stanley (1963) pointed out, these potentially restrictive factors are interaction effects and it is the interpretability or understanding of results due to these effects that is required.

In cases where construct definitions are of importance, external validity is concerned with experimental episode representativeness with respect to these definitions. One may ask how similar is this experimental situation to that one described within the construct definition and an answer will depend, as above, on the matters of similarity and interpretability. Regarding the former, one may examine the comprehensiveness of the experimental examination; that is, how carefully and completely it describes the individuals, treatments, and environments involved. With regard to the latter, one may examine the level of interpretability in terms of the construct of episode outcomes due to possible interaction effects.

Internal validity refers to the interpretability in terms of some theory or set of assumptions being considered of results due to experimental episode main effects. Thus internal validity helps to insure the internal consistency of these episodes, making it possible to understand them fully and increasing the probability of their accurate replication. In short, this validity allows the researcher to answer such questions as the one posed by Campbell and Stanley: "Did in fact the experimental treatments make a difference in this specific experimental instance?" These authors noted that "internal validity is the basic minimum without which any experiment is uninterpretable." In cases where construct definitions are of importance, therefore, internal validity provides for the sort of understanding of results in terms of the construct that allows for similar results to be obtained from a future experimental episode.

In cases involving construct definitions, therefore, internal and external validities together can be regarded as equivalent to the overall construct validities of experimental episodes. In the present study this equivalence is represented by α and λ; together they provide an assessment and make possible an evaluation of construct validity. Individually these parameters are to be regarded as coefficients of external and internal validity respectively. α, the parameter governing inductive generalization in Hintikka's system, is defined so that its value is a function of both experimental episode representativeness and the interpretability of episode outcomes due to interaction effects in terms of a construct. Thus this parameter determines how easily generalizations from these episodes may be made to the construct definition. λ, which for both Carnap and Hintikka governed inductive inference, is defined so that its value is a function of the interpretability of experimental episode outcomes due to main effects in terms of a construct. This parameter therefore determines the ease with which inference(s) may be made to future, similarly designed experimental episodes.

Assessing construct validity

Estimation of α and λ as coefficients of external and internal validity requires that both a construct definition and an experimental episode of interest be translated into logical symbols. In the case of the construct

definition this translation amounts to its being put in standard formalization or, even more simply, to its being characterized in terms of one-place predicates and sentential connectives. Translation of the episode into logical predicates is less straightforward In this case it is necessary to describe the ways in which particular test items or treatments can picture the person examined, to describe the persons themselves, the environment of the test/experimental situation, and so on. For these episodes any aspect that one has reason to believe will affect generalization to or inference in terms of the construct must be characterized logically. The two episodes to be considered here require only that test item descriptions and descriptions of a single person tested be translated into one-place predicates, so they serve as rather uncomplicated examples. Ten logical definitions (L-definitions) based on the translation of these episode aspects into logical predicates are fundamental elements in the defining formulas proposed later in this paper for α and λ.

A. Translation of experimental episodes into logical predicates

Given the assumption that each test item, for example, is designed to describe the person tested in a particular way, it is this description that must be logically characterized. Behaviors that the person tested must display in order to consider a test item (e.g., reading English, sitting quietly) are not characterized unless their performance is part of the description of the person the test was designed to produce. Consider an item from the objective test of academic achievement developed by Albert Myers (1964):

> 3. Have others (not your good friends) thought of you as one who "missed some of the fun" because you were so serious?

The individual who considers this question is asked whether or not "he or she sees him/herself being called serious by others." In the context of academic achievement the term "serious" can refer to the individual's "good work habits"--the fact that he/she "works hard or works well." The term "so" implies that the ability to work hard and well is something the individual demonstrates much of the time in school and perhaps at home also. Thus, "sees self called serious by others" can be read as "sees self as working hard much of the time."

This question also asks the individual to evaluate the attitudes of "others." If these "others" think he/she has missed fun and the individual understands this is their feeling but persists, then he/she must find value in hard work. The fact that others see him/her as missing fun on account of strict work habits can be read as a measure of the individual's own valuation of these habits. Within the context of academic achievement it also appears that, regardless of the values attached to the work itself, there is at least some emphasis placed on "working for a purpose or goal." It therefore seems that this question asks the individual whether he/she works hard, works for a purpose or goal, and values this purpose or goal. This is described schematically as follows:

a. Works hard/works to achieve goal/goal is valued
b. Works hard/wants to achieve goal/goal is valued
c. Works hard/wants to achieve goal/positive valuing
d. Works hard/wants/wants to achieve goal/positive valuing
e. Works hard/positive desire/wants to achieve goal/positive valuing

If a set of one-place predicates is chosen, this test item may be characterized logically as

3. W H N

where

W = works hard (at task)
H = associates positive affective states with goal attainment
N = desires to reach achievement goal

What the symbol W stands for may be further specified as

W = works hard (at task) =
C = uses overt or mental activity instrumentally to attain goal
E = while blocked by external obstacles to goal attainment
F = while blocked by internal obstacles to goal attainment

"External obstacles" can be thought of here as the subject matter itself, after-school commitments, or poor facilities for studying at home. "Internal

obstacles" can be seen as anything from the individual's dislike of the subject matter to his/her ineptness in working with numbers. Item 3 can now be expressed as WHN or, more specifically, as CEFHN.

In a way that is similar though not as complex, aspects of the person or persons tested that are thought to affect interpretation of test performance can be characterized by logical predicates. These aspects should be spelled out by the theory surrounding the construct although they usually are referenced implicitly. For example, while it appears that the universal quantifier and variables bound to it can be associated with each individual and complex predicate expression in assumptions surrounding the construct "need for achievement" (n Achievement), a more careful examination reveals that the universality of these assumptions is questionable.

Of the studies reviewed in Atkinson and Feather (1966), all but one reported experiments involving male subjects only.[3] In addition, of these studies involving males, only three included noncollege males and one of these three was carried out in Wales. The fact that males predominated in the studies suggests these assumptions should be restricted to males only (see Pringle, 1973 for a discussion of problems associated with the measurement of n Achievement in females). Additionally, since the majority of these studies were undertaken with college males, there is reason to believe that "males" is too broad a category. College males are usually between 19 and 23 years old and are likely to be (this would have been even more true in the late 1950's and early 1960's) White or Anglo. Sex, age, and ethnic background thus seem to be variables which can affect ease of inference or generalization from experimental episodes assessing n Achievement. Predicates denoting these variables are therefore included in the logical characterization of this construct definition and the person tested with the Myers test and the one developed by McClelland et al is described in terms of them. For the present analysis the predicates R, S, and T will denote "is male," "is 19-23 years of age," and "is White/Anglo" respectively.

---

[3] These studies are: French and Thomas (1958), Litwin (1958), Winterbottom (1958), Feather (1959), Atkinson and Litwin (1960), Mahone (1960), Atkinson and O'Connor (1961), Feather (1961), Smith (1961), Brody (1963), Feather (1963), Littig (1963), Morgan (1964), Feather (1965), and Moulton (1965).

B. Ten L-definitions

L-definitions are so called because they describe logical or structural properties of experimental episodes. For example, test items that are logically characterized in terms of what they are designed to elicit from individuals tested can be viewed in a strictly logical or structural way. Consider a hypothetical three-item test characterized logically:

Item 1. A
Item 2. B
Item 3. C

where A = can write his/her name phonetically
B = can read a simple two-word sentence
C = can identify the capital of Egypt

Notice that the test items themselves are not described by the predicates A, B, and C. These items may be in various forms such as written tasks, oral responses, selection of flashcards, and so on. What is described is the "examining purpose" of each item; that is, A describes what item 1 is designed to tell about the individual tested, likewise for B and C. Descriptions of this sort are therefore referred to as L-structures rather than as test descriptions.

From these three predicate symbols one L-item is formed, viz. A & B & C or ABC. An L-item represents one complete examination of the individual tested. A "complete examination" in this sense consists of a single evaluation described by a set of predicates such that this set includes one and only one occurrence of each and every distinct predicate symbol representing test items. Consider the following eight examples:

(1) Item 1. A
Item 2. B
Item 3. C

(2) Item 1. ABC
Item 2. A
Item 3. BC

(3) Item 1. A
Item 2. BC
Item 3. AC

(4) Item 1. ABC
Item 2. CA
Item 3. B

(5) Item 1. ABC
Item 2. ABC
Item 3. ABC

(6) Item 1. ABC
Item 2. A
Item 3. B

(7) Item 1. A
Item 2. B
Item 3. C
Item 4. A
Item 5. B
Item 6. C

(8) Item 1. AB
Item 2. B
Item 3. C
Item 4. AC

In (1) there is but <u>one</u> L-item made up of test items 1, 2, and 3. In (2) there are <u>eight</u> L-items defined by all logically possible combinations of the predicates from items 1 and 2-3. In (3) there are <u>four</u> L-items formed from the following test item combinations: 1-2, 1-2-3 (with the predicate A from item 3 and the predicate C from item 2 omitted), 2-3 (with the predicate C from item 2 omitted), 2-3 (with the predicate C from item 3 omitted). In (4) there are <u>eight</u> L-items formed from all logically possible combinations of the predicates from items 1 and 2-3. In (5) there are <u>twenty-seven</u> L-items--3^3 combinations derived from all logically possible combinations of the predicates in test items 1, 2. and 3. In (6) there are <u>four</u> L-items. In (7) there are <u>eight</u> L-items while in (8) there are also <u>eight</u> L-items. In short, when each logical predicate representing a test item aspect can be used to characterize a person tested without ambiguity, L-items are formed from all logically possible combinations of these predicates.

L-items are defined and their total number derived once test items are characterized logically. In terms of L-items <u>L-item descriptions</u> or "results" from L-item examinations are derived. For example, from the one L-item ABC, eight L-item descriptions can be formed by permuting these predicates with respect to negation (ABC, ~ABC, A~BC, AB~C, ~A~BC, ~AB~C, A~B~C, ~A~B~C). Notice, however, that the L-item ABC and the L-item description ABC are <u>not</u> the same.

An analogy may be helpful in making the distinction between an L-item and an L-item description. Think of an L-item as a roll of film and the L-item descriptions derived from the L-item as negatives on the film. The more predicate symbols that make up the L-item, the longer is the roll of film--but it is still <u>one</u> roll of film. The longer the roll of film, the more negatives there are on it and the more pictures that can be produced. Suppose now that the entire roll of film is used to take pictures of the

same individual. The roll of film then contains several unique descriptions of this individual. If the film is developed and only that picture formed from the clearest negative is kept, then this negative is analogous to the L-item description determined for some individual by test examination. Increasing the number of L-items has the same effect as increasing the number of film rolls used, whereas an increase in the number of possible L-item descriptions depends on how "long" is each roll used. As for determining which L-item description most adequately characterizes the individual, this depends, in testing as well as in picture-taking, on which description is the clearest one produced.

Similar to the distinction between L-items and L-item descriptions is the one between L-persons and L-person descriptions. In this case the basic set of predicates is determined by those characteristics of the person tested which are believed to affect generalization or inference from the experimental episode. These characteristics can be as general as sex, race, or age, or as specific as scores on maze tracing and perceptual speed tests or a self-reported description of interests. L-persons provide a single, complete format for examination of individuals to be tested. When particular person-characteristics are determined they form an L-person description. As might be expected, L-items and L-persons and the descriptions formed from them are the basic elements of the two-element experimental episodes (episodes made up of test and tested subject) dealt with explicitly in this study. These descriptions are regarded as results due to the main effects of these test episodes when considered separately and as results due to the interaction effects of these episodes when combined to form L-item × L-person descriptions.

For both L-items and L-persons the number of predicates included in their formats is important. L-comprehensiveness refers to this format width. When considered along with the "response mode" of a test (e.g., true-false, multiple choice, self-written paragraph) this aspect provides a way of evaluating the potential clarity of results. Analysis of the Myers test to be presented later shows that its response mode is inefficient due to the high level of L-comprehensiveness of its L-item format.

Two logical definitions that are closely related to L-comprehensiveness are L-length and L-number. L-length refers to the total number (N) of L-items described by a particular test. L-number refers to the total number (Np) of L-persons or, more simply, of persons tested. In the present study only two-element episodes are considered but the definitions presented are easily extended to a greater number of elements. For example, L-place, L-place description, the L-comprehensiveness of an L-place, and the L-number (Npl) of L-places are all easily incorporated.

The final L-definition to consider is L-parallelism, which refers to the degree of definability of the construct in terms of experimental episode-produced descriptions. In particular, L-parallelism refers both to the definability of the construct on results from episode main effects--L-item descriptions and L-person descriptions considered separately, and to the definability of the construct on results from episode interaction effects--L-item descriptions and L-person descriptions considered together (L-item × L-person descriptions). For example, if AB is an L-item and RS is an L-person, L-parallelism refers in this case to how well the construct is defined by the following results: AB, ~AB, A~B, ~A~B; RS, ~RS, R~S, ~R~S; and ABRS, ~ABRS, A~BRS, AB~RS,..., ~A~B~R~S.

C. Formulas for defining α and λ

Using these L-definitions it is possible to sketch how the coefficients α and λ are estimated. α, the coefficient of external validity, must be defined as a function of the following two factors: (1) the overall representativeness of the experimental episode with respect to the universe of generalization and (2) the overall interpretability of episode outcomes due to interaction effects in terms of this universe. (1) is defined in this study by the product of the numbers of different kinds of episode observations weighted by the degree of interpretability of derived descriptions. It is the weighted product of these observations rather than their sum that is considered because each sort of observation is absolutely necessary for evaluating representativeness. For example, in the case of person-test episodes, observations of both persons and tests are required or the external validity of the episode is indeterminable. The weighting of this product is necessary in order that it is the product of episode observations relevant to the construct-defined universe that is referred to. (2) is defined

by the L-parallelism of episode results due to interaction effects with respect to the universe of generalization. λ, the coefficient of internal validity, is defined by the overall interpretability of episode outcomes due to main effects. For λ, the matter of episode representativeness is not such an important factor. This is due to the different inductive disposition which this validity describes. Internal validity is concerned primarily with episode replicability and thus it is important only whether results due to main effects in an episode are well understood.

The following formulas are therefore used to estimate values for α and λ:

$$\alpha = \frac{K_{(txpxi)} - K_{(txpxi)c}}{N \cdot Np \cdot Ni\left(\frac{K_{(txpxi)c}}{K_{(txpxi)}}\right)}$$

$$\lambda = ((K_t - K_{tc}) + (K_p - K_{pc}) + (K_i - K_{ic})$$

where

K_t = total number of L-item descriptions

K_{tc} = total number of theory-defined L-item descriptions

K_p = total number of L-person descriptions

K_{pc} = total number of theory-defined L-person descriptions

K_i = total number of other relevant experimental episode L-descriptions

K_{ic} = total number of theory-defined, other relevant experimental episode L-descriptions

$K_{(txpxi)}$ = total number of L-item x L-person x other relevant episode L-descriptions

$K_{(txpxi)c}$ = total number of L-item x L-person x other relevant episode L-descriptions that are theory-defined

N = L-length (total number of L-items)

N_p = L-number of persons (total number of persons tested)

N_i = L-number of observations provided by other relevant episode elements

For the two-element (person-test) experimental episodes considered here, these formulas may be rewritten as

$$\alpha = \frac{K_{(txp)} - K_{(txp)c}}{N \cdot Np\left(\frac{K_{(txp)c}}{K_{(txp)}}\right)}$$

$$\lambda = ((K_t - K_{tc}) + (K_p - K_{pc}))$$

where the symbols have meanings as above

The coefficients α and λ thus take on their largest values when no episode-produced descriptions are interpretable in terms of the construct. In this case, the numerator of α will be large but, since the ratio $\frac{K_{(txpxi)c}}{K_{(txpxi)}}$ has a numerator of zero, the denominator of α will be zero thus causing $\alpha \longrightarrow \infty$. For λ, total uninterpretability means that the differences $K_t - K_{tc}$, $K_p - K_{pc}$, and $K_i - K_{ic}$ will take on their largest values signaling the need for extreme caution when making inferences to future test episodes. In terms of these formulas the external and internal validities of test episodes may be assessed at certain time points. When the universe of interest is construct-defined these formulas thus provide an overall assessment of construct validity. In either case these formulas represent formally the "cross-purposes" of internal and external validities. As Campbell and Stanley noted, the researcher must aim for "that maximum similarity of experiments to the conditions of application which is compatible with internal validity" (p. 188). Thus, in the assessment of these validities it is interesting to compare the overall representativeness of the episodes considered (denoted by the relation of $N \cdot Np \cdot Ni\frac{K_{(txpxi)c}}{K_{(txpxi)}}$ to $K_{txpxi} - K_{(txpxi)c}$) to the interpretability of their possible outcomes due to main effects.

D. Example analyses

The methods and definitions presented in this section can be put to use to determine α and λ values for two experimental episodes designed to assess

the construct n Achievement. The construct definition presented is a simplified version of the one presented by Atkinson (1957) and the tests considered are the "focused" Thematic Apperception Test (TAT) developed by McClelland et al (1953) and the twelve-item objective test of academic achievement developed by Myers.

1. Definitions and assumptions related to *n* Achievement

a. Definitions

A = n Achievement (this may be considered as equivalent to the situation where the motive to succeed is stronger than the motive to avoid failure, i.e. Ms > Maf

N = desire to reach some goal

H = positive affective states associated with goal attainment

K = negative affective states associated with failure to attain goal

E = external obstacle to goal attainment

F = internal obstacle to goal attainment

C = overt or mental activity used instrumentally for goal attainment

G = personal forces or aids which help goal attainment

((EvF) & G) = Po (where Po = "objective probability of success on some task is known," and it is presumed that some combination of obstacles to success and means available for overcoming these obstacles can result in derivation of such an objective measure).

D = anticipation of goal attainment

J = anticipation of failure to attain goal

((EvF) & G) → D & J = Ps = .5 = Ps (where Ps = subjective probability of success, and it is presumed that some combination of obstacles and means for overcoming these obstacles implying that equal anticipation of success or of failure is warranted will result in equally valued subjective estimates of success or failure)

((EvF) & G) → D = Ps > .5 (it is presumed that where some combination of obstacles and means for their resolution warrant anticipation of success, subjective probability estimates of success will, usually, be higher than those for failure)

((EvF) & G) → J = Ps < .5 (it is presumed that where some combination of obstacles and means for their resolution warrant anticipation of failure, subjective probability estimates of success will, usually, be lower than those for failure)

b. Assumptions

1. $A \rightarrow (N \ \& \ H)$
{n Achievement leads to the establishment of goal-directed desires and to positive affective states being associated with success at reaching such goals}

2. $\sim A \rightarrow (\sim N \ \& \ K)$
{Low n Achievement leads to the disestablishment of goal-directed desires and to negative affective states being associated with failure to reach such goals}

3. $A \ \& \ (((EvF) \ \& \ G) \rightarrow D \ \& \ J) \rightarrow C$
{High n Achievement + Ps = .5 leads to acceptance of task}

4. $\sim A \ \& \ (((EvF) \ \& \ G) \rightarrow D \ \& \ J) \rightarrow \sim C$
{Low n Achievement + Ps = .5 leads to non-acceptance of task}

Conjoining to these assumptions the predicates R = "is male," S = "is 19-23 years of age," and T = "is White/Anglo," 2^{13}=8192 a-constituents can be derived, compared to the four assumptions, and used to identify those episode-produced a-constituents (L-item, L-person, and L-item × L-person descriptions) in terms of which n Achievement may be defined empirically.

2. Logical characterization of two experimental episodes

The "focused" TAT consists of four pictures for which the subject is asked to make up a story. Scoring is based on a clinician's rating of the story according to eleven achievement-related criteria. These criteria are the following:

AI = shows concern for competition with some standard
N = wants to reach an achievement goal
I = overt or mental activity instrumental for goal attainment
Ga = anticipates goal attainment
Bp = a block or obstacle within the person
Bw = a block or obstacle external to the person
Nup = personal or nurturing forces able to assist in goal attainment
G = evidence of positive affective states associated with goal attainment
-Ga = anticipates failure

-G = evidence of negative affective states associated with failure to reach goal

AchTh = achievement imagery elaborated into a plot or theme

Depending on the story told, the subject is described according to some or all of these test criteria.

In the present analysis, AI (concern for competition) and AchTh (achievement imagery elaborated into a plot or theme) are regarded as measures of the intensity of n Achievement rather than of n Achievement itself. For this reason these points are not included in analyses to follow. The remaining nine descriptors are characterized in the following way in order that they may be compared to theoretically-defined elements.

Test Descriptor	System Descriptor
N	N
I	C
Ga	D
Bp	F
Bw	E
Nup	G
G	H
-Ga	J
-G	K

Given this set of descriptors, the four test items for the focused TAT may be characterized logically as follows:

Item 1: L-item = NCDFEGHJK

Item 2: L-item = NCDFEGHJK

Item 3: L-item = NCDFEGHJK

Item 4: L-item = NCDFEGHJK

The L-item structure of the test is NCDFEGHJK while N (the number of observations provided by the test) is equal to 4^9=262,144. Since the L-item format is made up of nine predicates, there are 2^9=512 test-produced L-item descriptions possible. If the episode in which the focused TAT is administered involves a subject who is male, between 19 and 23 years of age, and white/Anglo then the L-person description to be considered is RST.

If, on the other hand, the test is administered to several subjects, or alternatively it is suggested that the test is, in principle, administerable to anyone, then RST is the L-person format such that there are $2^3=8$ possible test episode L-person descriptions. In this way, depending on the number of individuals tested, test episode external and internal validities can be computed for any group, subgroup, or particular individual of interest.

The items on the Myers scale typically ask the subject "whether he works harder than most to achieve academically" (e.g., "to make the honor roll or merit list"). The following system descriptors will be used both to focus the analysis of these items and to make the results comparable to theoretically specified elements:

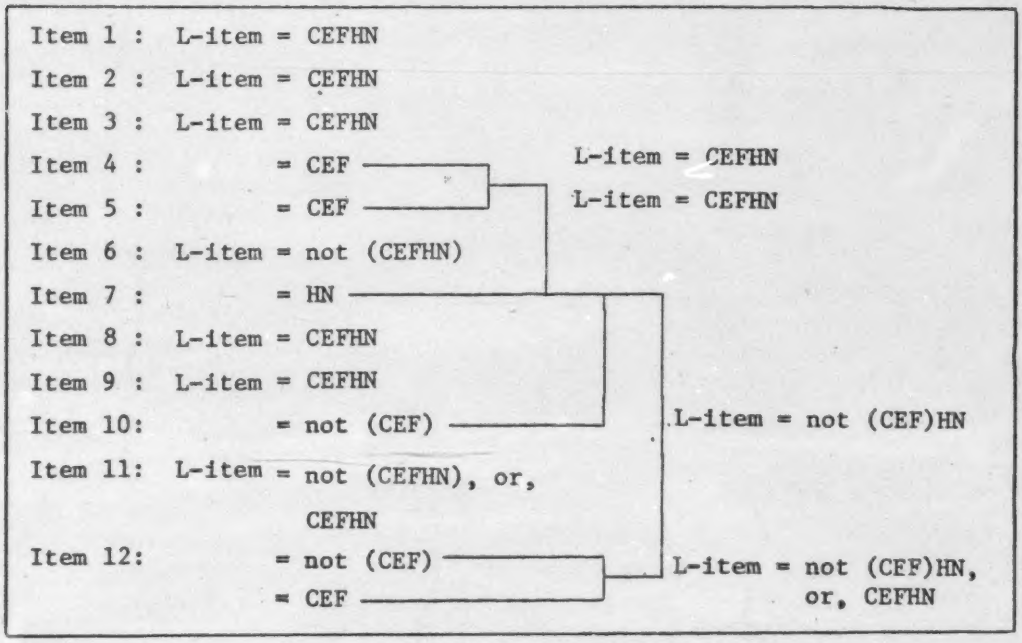
Test Descriptors	System Descriptors
Works hard (at task)	W = CEF
Evidence of positive affective states associated with goal attainment	H
Wants to reach achievement goal	N
Overt or mental activity instrumental to goal attainment	C
Anticipates goal attainment	D
Obstacle within person	F
Obstacle external to person	E
Personal forces able to assist in goal attainment	G
Anticipates failure	J
Evidence of negative affective states associated with failure to reach goal	K

Translation of each item on this scale into the system descriptors listed is accomplished by the sort of analysis summarized in Table 2. Results from analysis of the complete test are illustrated in Table 3.

TABLE 2. Methods of Item Analysis and Logical Characterization Applied to Myers' Test of Academic Achievement

Item	1. When you know there are going to be one or two questions on an exam from outside reading assignments, do you always read *all* the material?
Analysis	___ does maximum amount of work to do well on exam ___ works hard to achieve (high) grade on exam ___ works hard to achieve goal ___ works hard/wants to achieve goal ___ works hard/wants/ /wants to achieve goal/ ___ works hard/positive desire/wants to achieve goal
Translation	___ W H N ___ CEF H N

TABLE 3. Results from Analysis of Myers' Test of Academic Achievement


Item 1 : L-item = CEFHN
Item 2 : L-item = CEFHN
Item 3 : L-item = CEFHN
Item 4 : = CEF
Item 5 : = CEF
L-item = CEFHN
L-item = CEFHN
Item 6 : L-item = not (CEFHN)
Item 7 : = HN
Item 8 : L-item = CEFHN
Item 9 : L-item = CEFHN
Item 10: = not (CEF)
L-item = not (CEF)HN
Item 11: L-item = not (CEFHN), or, CEFHN
Item 12: = not (CEF)
= CEF
L-item = not (CEF)HN, or, CEFHN

Since each L-item consists of 5 predicates it is easily seen that $2^5=32$ L-item descriptions are possible.

The Myers test is, however, an example of a test which suffers from ambiguity due to a too L-comprehensive L-item structure for its response mode. Unlike the TAT, which has a response mode capable of producing clear descriptions for subjects tested, the Myers test can produce such a description only if CEFHN is obtained. In the other four cases where L-items are formed from negations of predicate groupings or from predicate combinations across test items (viz. where the L-item description is either :~(CEFHN), ~(HN)CEF, HN~(CEF), or ~(HN)~(CEF)) the result is unclear. This ambiguity also restricts the number of possible L-item examinations since it is impossible to determine clearly the values of C, E, F, H, or N for any one L-item embedded in a particular test item. Thus, each L-item noted in Table 3 can produce but one test item-specific L-item examination. Consider the difference between the TAT with $4^9=262,144$ possible L-item examinations and the Myers test with only 11! Were the Myers test without ambiguity this scale would have $11^5=161,051$ L-items (since the results of individual rather than only grouped "predicate" examinations could be combined across test items).

3. α and λ estimates

Since the number of observational predicate combinations which explicitly define n Achievement is determinable by the procedure outlined above and the total N's and predicate formats for each of the two test episodes have been characterized, values for the coefficients α and λ can be derived. (In the two-element episodes considered here Np = 1, characterized by the L-person description RST.)

TAT:

$$\alpha = \left(\frac{K_{(txp)} - K_{(txp)c}}{N \cdot Np \; \frac{K_{(txp)c}}{K_{(txp)}}} \right) = \left(\frac{512 - 163}{262,144(.3183)} \right) = \frac{349}{83,440.43} = .0042$$

$$\lambda = (K_t - K_{tc}) + (K_p - K_{pc}) = (512 - 163) + (1 - 1) = 349$$

Myers:

$$\alpha = \left(\frac{K_{(txp)} - K_{(txp)c}}{N \cdot N_p \; \frac{K_{(txp)c}}{K_{(txp)}}} \right) = \left(\frac{32 - 21}{11(.6561)} \right) = \frac{11}{7.22} = \underline{1.5235}$$

$$\lambda = (K_t - K_{tc}) + (K_p - K_{pc}) = (32 - 21) + (1 - 1) = \underline{11}$$

α for the TAT is much smaller than α for the Myers test because of the greater L-comprehensiveness of its L-items and the clarity with which their component predicates can be understood. λ for the TAT is much larger than λ for the Myers test because the L-comprehensiveness of the TAT is large while the number of construct-defining descriptions is small. This is to be expected as the greater L-comprehensiveness of the TAT should make it more unlikely that the same results will be obtained for the person tested in a future experimental episode.

Evaluating construct validity

The construct validity of an experimental episode must be evaluated both in terms of its inductive potential and in terms of its informativeness. Since α and λ values provide an assessment of inductive potential, the TAT and Myers test episodes can be evaluated on this aspect given the results above. It is evident that the TAT episode is a superior one for generalizing to the construct n Achievement while the Myers test episode appears more suited for making inferences to similarly designed episodes. Assessing the informativeness of these episodes--in particular their person and construct informativeness--requires that probability values and information measures associated with selected a-constituents and constituents (L-structures) from these episodes be derived. For this evaluation-aspect α and λ values are considered only indirectly.

A. Person and construct informativeness of the TAT and Myers episodes

Substituting values for α, λ, and $N \cdot Np = n$ from the above analyses into the formulas presented earlier and specifying values for c, n_j, and w, the

a priori probabilities for a selection of constituents (Cw's), a posteriori probabilities for these constituents (Cw/e's), and the a posteriori probability for a selected a-constituent (h/e) can be derived. Tables 4 and 5 show these probability values and information measures derived from them for the TAT and the Myers episodes. In reading these tables three "relations" deserve special attention. These relations are those between: the width of a constituent (w) and the comprehensiveness of available evidence (e), a priori/a posteriori probability values and a priori/a posteriori information measures, and a priori measures of information and a posteriori probabilities.

A constituent is a generalization; in the present case it can be considered as an L-score advanced on evidence provided by one or the other of the episodes being considered. The width of a constituent thus represents how encompassing is this generalization. For this reason it is to be expected that the most probable generalizations advanced on the basis of episode results reported in Tables 4 and 5 are those with widths "equal" to the variety of evidence at hand. In other words, the most probable constituent in each case is that one which contains all and only those a-constituents (L-item × L-person descriptions) included in the evidence (c). For the TAT this constituent is Cw where w=c=8 while for the Myers test episode it is Cw where w=c=1. The wider the generalization given the same body of evidence the less is its probability. This explains why p(Cw/e) values are greatest for the constituent Cc. Examining this value for both tests, it is clear that the TAT episode offers much more in terms of person informativeness.

Since the relation between probability values and measures of information is an inverse one, the researcher must consider how well balanced are these two aspects of experimental episodes before and after data are collected. For example, it is useful to ascertain whether a construct informative statement one would like to make also is likely to be informative with respect to the person(s) tested. Thus the values of p(Cw), p(Cw/e), and p(h/e) within each of the cells of Tables 4 and 5 should be compared to their respective cont and inf measures in these cells. Evidently, neither the TAT nor the Myers test episode fares especially well in establishing this balance.

TABLE 4. Logical Probability Values and Measures of Semantic Informativeness for the TAT

α=.0042, λ=349, K=512, c=8, n=262,144, n_1=50,000, n_2=50,000, n_3=50,000, n_4=50,000, n_5=30,000, n_6=30,000, n_7=2,000, n_8=144		
w=8: p(Cw) = 0 cont(Cw) = 1 inf(Cw) = 354.91	p(Cw/e) = 1 $cont_{cond}$(Cw/e) = 0 inf_{cond}(Cw/e) = 0	p(h/e) = .19 $cont_{cond}$(h/e) = .81 inf_{cond}(h/e) = .7212
w=30: p(Cw) = 0 cont(Cw) = 1 inf(Cw) = 354.90	p(Cw/e) = 0 $cont_{cond}$(Cw/e) = 1 inf_{cond}(Cw/e) = 1417.5	p(h/e) = .19 $cont_{cond}$(h/e) = .81 inf_{cond}(h/e) = .7212
w=65: p(Cw) = 0 cont(Cw) = 1 inf(Cw) = 354.90	p(Cw/e) = 0 $cont_{cond}$(Cw/e) = 1 inf_{cond}(Cw/e) = 1777.6	p(h/e) = .19 $cont_{cond}$(h/e) = .81 inf_{cond}(h/e) = .7212

TABLE 5. Logical Probability Values and Measures of Semantic Informativeness for the Myers Test

α = 1.5235, λ = 11, K = 32, C = 1, n = 11, n_1 = 11		
w=1: $p(Cw) = .16 \times 10^{-10}$ cont(Cw) = .9999 inf(Cw) = 24.859	$p(Cw/e) = .257 \times 10^{-6}$ $cont_{cond}$(Cw/e) = .9999 inf_{cond}(Cw/e) = 15.173	p(h/e) = .53 $cont_{cond}$(h/e) = .47 inf_{cond}(h/e) = .2757
w=14: $p(Cw) = .17 \times 10^{-9}$ cont(Cw) = .9999 inf(Cw) = 22.493	$p(Cw/e) = .5064 \times 10^{-9}$ $cont_{cond}$(Cw/e) = .9999 inf_{cond}(Cw/e) = 21.404	p(h/e) = .53 $cont_{cond}$(h/e) = .47 inf_{cond}(h/e) = .2757
w=32: $p(Cw) = .61 \times 10^{-9}$ cont(Cw) = .9999 inf(Cw) = 21.217	$p(Cw/e) = .3602 \times 10^{-9}$ $cont_{cond}$(Cw/e) = .9999 inf_{cond}(Cw/e) = 21.744	p(h/e) = .53 $cont_{cond}$(h/e) = .47 inf_{cond}(h/e) = .2757

A more important relation between probability values of constituents and their informativeness, however, is the one between a priori measures of information and a posteriori probabilities. In their 1966 paper entitled "Semantic Information and Inductive Logic," Hintikka and Pietarinen stated that "there is no need for absolute information and for a posteriori probability to be related inversely in the same way as absolute information and a priori probability are inversely related" (p. 106). These authors argued that science may in fact be best served by the search for hypotheses and generalizations that are most likely to be informative and are, at the same time, most easily confirmed. Carrying this further, experimental episodes are thus to be preferred which allow very informative generalizations to be advanced that are easily confirmed by evidence. Evaluation of this aspect of the above episodes is accomplished by comparing a priori inf and cont measures for the constituents presented in each Table with their a posteriori probability values. In this case the TAT episode is clearly superior. When the width of a generalization based on this experimental episode is as encompassing as the variety of evidence available (when w=c=8), this generalization is both maximally informative a priori (cont(Cw)=1 and inf(Cw)=354.91) and most easily confirmed (p(Cw/e)=1). This same generalization (w=c=1) advanced on the Myers test episode is neither as informative a priori (cont(Cw)=.9999 and inf(Cw)=24.859) nor as easily confirmed ($p(Cw/e)=.257\times10^{-6}$).

B. Resolving the Popper-Carnap controversy: Significance for research

The way of resolving Karl Popper's concern for the informativeness of hypotheses with Rudolf Carnap's concern for their confirmability that was suggested by Hintikka and Pietarinen and is described in the preceding paragraph is given practical significance by the present application of inductive methods to the analysis of experimental episodes. The defining formulas for α and λ proposed above enable derived probability values to be interpreted as measures of person informativeness in terms of a construct of interest. These formulas also allow semantic information measures to be regarded as indicators of the potential construct informativeness of the episode being considered. Since researchers involved in the process of construct validation usually are concerned both for describing persons tested accurately in terms of the construct and for strengthening/extending a considered construct definition, these two types of informativeness must be

taken into consideration when experimental episodes are designed and analyzed. The concerns of Karl Popper and Rudolf Carnap are therefore the concerns of researchers, and application of inductive methods to assessing and evaluating construct validity can help understand the importance of resolving these concerns for the practice of research.

References

Atkinson, J. W. Motivational determinants of risk-taking behavior. Psychol. Review, 1957, 64, 359-372. (References here are made to Reprint in J. W. Atkinson & N. T. Feather (Eds.), A theory of achievement motivation. New York: Wiley, 1966.

Atkinson, J. W., & Feather, N. T. (Eds.) A theory of achievement motivation. New York: Wiley, 1966, 11-29.

Atkinson, J. W., & Litwin, G. H. Achievement motive and test anxiety conceived as motive to approach success and motive to avoid failure. J. Abnormal Soc. Psychology, 1960, 60, 52-63.

Atkinson, J. W., & O'Conner, P. Effects of ability grouping in schools related to individual differences in achievement-related motivation (Project 1283). Cooperative Research Program of the Office of Education, U. S. Department of Health, Education, and Welfare, 1961.

Bar-Hillel, Y. Semantic information and its measures. Transactions of the Tenth Conference on Cybernetics, New York, 1952.

Bar-Hillel, Y. Language and information. Reading, Mass.: Addison-Wesley, 1964.

Brody, N. N achievement, test anxiety, and subjective probability of success in risk taking behavior. Journal of Abnormal Social Psychology, 1963, 66, 413-418.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. Bull., 1959, 56, 81-105.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. Chapter 5 in N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 171-246.

Carnap, R. Logical foundations of probability. Chicago: University of Chicago Press, 1950.

Carnap, R. Inductive logic and science. Paper read at the Symposium on Philosophy of Science, joint session of the Amer. Phys. Soc. and the Philosophy of Science Association, Chicago, 1951. (Also in Proceedings of the American Academy of Arts and Sciences, 1953, 80, No. 3.)

Carnap, R., & Bar-Hillel, Y. An outline of a theory of semantic information (Technical Report No. 247). Cambridge: Research Laboratory of Electronics, 1952.

Cronbach, L. A generalized psychometric theory based on information measure: A preliminary report. Urbana: Bureau of Research and Service, College of Education, University of Illinois, 1952. (Mimeographed)

Cronbach, L. A consideration of information theory and utility theory as tools for psychometric problems. Urbana: Bureau of Research and Service, College of Education, University of Illinois, 1953. (Mimeographed)

Cronbach, L., & Meehl, P. Construct validity in psychological tests. Psychol. Bull., 1955, 52, 281-302. (References here are made to Reprint in W. Mehrens & R. Ebel (Eds.), Principles of educational and psychological measurement: A book of selected readings. Chicago: Rand McNally, 1967, 243-270.

Feather, N. T. Subjective probability and decision under uncertainty. Psychol. Review, 1959, 66, 150-164.

Feather, N. T. The relationship of persistence at a task to expectation of success and achievement related motives. J. Abnorm. Soc. Psychol., 1961, 63, 552-561.

Feather, N. T. Persistence at a difficult task with alternative task of intermediate difficulty. J. Abnorm. Soc. Psychol., 1963, 66, 604-609.

Feather, N. T. The relationship of expectation of success to need achievement and test anxiety. J. Pers. Soc. Psychol., 1965, 1, 118-126.

French, E., & Thomas, F. The relation of achievement motivation to problem-solving effectiveness. J. Abnorm. Soc. Psychol., 1958, 56, 46-48.

Hintikka, K. J. Distributive normal forms in the calculus of predicates. Acta Phil. Fennica, 1953, 66.

Hintikka, K. J. Distributive normal forms in first-order logic. In J. N. Crossley and M. A. E. Dummett (Eds.), Formal systems and recursive functions: Proceedings of the eighth logic colloquium. Amsterdam: North-Holland Publishing Company, 1963.

Hintikka, K. J. A two-dimensional continuum of inductive methods. In K. J. Hintikka & P. Suppes (Eds.), Aspects of inductive logic. Amsterdam: North-Holland Publishing Company, 1966, 113-132.

Hintikka, K. J. The varieties of information and scientific explanation. In B. van Rootselaar & J. Staal (Eds.), Logic, methodology, and philosophy of science III: Proceedings of the 1967 international congress. Amsterdam: North-Holland Publishing Company, 1967, 311-331.

Hintikka, K. J., & Pietarinen, J. Semantic information and inductive logic. In K. J. Hintikka & P. Suppes (Eds.), Aspects of inductive logic. Amsterdam: North-Holland Publishing, 1966, 96-112.

Hintikka, K. J., & Tuomela, R. A general theory of auxiliary concepts. In K. J. Hintikka & P. Suppes (Eds.), Information and inference. Dordrecht: D. Reidel Publishing, 1970, 298-330.

Littig, L. Effects of motivation on probability preferences. J. of Personality, 1963, 31, 417-427.

Litwin, G. Achievement motivation, expectancy of success, and risk-taking behavior. Chapter 7 in J. Atkinson & N. Feather (Eds.), A theory of achievement motivation. New York: Wiley, 1966.

Lord, F., & Novick, M. Statistical theories of mental test scores. Reading: Addison-Wesley, 1968.

Mahone, C. Fear of failure and unrealistic vocational aspiration. J. Abnorm. Soc. Psychol., 1960, 60, 253-261.

McClelland, D., Atkinson, J., Clark, R., & Lowell, E. The achievement motive. New York: Appleton-Century-Crofts, 1953.

Morgan, J. The achievement motive and economic behavior. In Economic Development and Cultural Change, 1964, 12, 243-267.

Moulton, R. Effects of success and failure on level of aspiration as related to achievement motives. J. of Personality and Soc. Psychol., 1964, 1, 399-406.

Myers, A. Risk-taking and academic success and their relation to an objective measure of achievement motivation. N ONR 2959(00). Princeton: Educational Testing Service, 1964.

Popper, K. Degree of Confirmation. British Journal for the Philosophy of Science, 5 (1954), 143-149.

Pringle, M. Using achievement motivation training with females. In R. Smith & G. Walz (Eds.), Developing students' potentials. Washington, D. C.: Educational Resources Division, Capitol Publishing Co., 1972, 18-22.

Shannon, C., & Weaver, W. The mathematical theory of communication. Urbana: Univ. of Illinois Press, 1949.

Smith, C. The influence of testing conditions on need for achievement scores and their relationship to performance scores. Chapter 18 in J. Atkinson & N. Feather (Eds.), A theory of achievement motivation. New York: Wiley, 1961, 277-298.

Snow, R. Representative and quasi-representative designs for research on teaching. RER, 1974, 44, No. 3, 265-291.

Tuomela, R. On construct validity of tests. Ajatus, 1970, 32, 47-66.

Winterbottom, M. The relation of need for achievement to learning experiences in independence and mastery. In J. Atkinson (Ed.), Motives in fantasy, action, and society. Princeton: Van Nostrand, 1958.